Steven Verstockt, Kenzo Milleville, Dilawar Ali, Francisco Porras-Bernardez, Georg Gartner, Nico Van de Weghe

EURECA - EUropean Region Enrichment in City Archives and collections

Keywords: *location based services, spatio-temporal mapping, social media analysis, natural language processing, handwritten text recognition, machine learning.*

Summary: The interdisciplinary EURECA project follows the trends towards location based services and personalized/contextualized content, and investigates them in the context of cultural heritage collections of European city archives. EURECA focuses on finding traces of European regions that have shaped the cities in which we live today and develops tools to easily explore them when visiting a city. The spatio-temporal metadata that is automatically generated by our tools can be used as input to perform new fundamental research and applied studies, but also to facilitate the exploitation of the collections to a broader public and attract new groups of cultural heritage consumers. LBS that run on top of our enrichments, for example, will allow tourists to explore the traces of a specific European region (e.g. Austria) in the city (e.g. Ghent) and show them the collection items at their corresponding point of interest (POI) using their mobile device. These connections that link Austria to Ghent (for example) can be rather diverse, such as architectural traces, art, place and street names, and memories of foreign academics. Different media-types, such as handwritten student/foreigner registers, pictures, newspapers, and wiki pages, are investigated. Natural language processing (NLP), computer vision and semantic intelligence techniques are the computational tools that are combined to automatically enrich the media items and link them to a particular Region of Origin (RoO). Furthermore, to get an idea about what cultural heritage items are popular by visitors from a particular region of origin, we also query social media platforms and generate RoO heatmaps of the cultural heritage points of interest (POIs) of today. These heatmaps can be compared across different regions or over different periods in time using standard and advanced GIS-techniques.

1. Introduction

Different historical, architectural, economic, political and cultural reasons have shaped the cities in which we live today. The main goal of the EURECA (EUropean Region Enrichment in City Archives and collections) project is to use input from each of these domains to reveal the cultural heritage items that can be linked to specific European regions/origins (~traces). The success of guiding tours focusing on such traces confirms that tourists like finding such traces and their also exist several websites that collect such links. “El rincón de sele”[1], for example, is a Spanish website on which you can perform region-based queries for Spanish traces, such as finding the Karel V points of interest (POIs) in Ghent. The traces of a region that can be found all over Europe, encompass the influence of European history of a single region in all regions within Europe - in which Europe is more than the sum of its states.

[1] https://www.elrincondesele.com/

The main goal of EURECA is to semi-automatically unearth European regions/origin traces in city archives and collections based on computational and crowdsourced (meta)data enrichment techniques. Our tools/applications will allow different types of end users (e.g. archivists, guides, researchers or European tourists in Ghent) to easily find the locations in the city where they can find cultural heritage connections to a particular European region/origin. These connections that link Austria to Ghent (for example) can be rather diverse, such as architectural traces (e.g. Hotel Falligan and the Austrian Military headquarters at the Handelsbeurs), historical facts (such as the introduction of house numbers by Jozef II), cultural events (e.g. Mozart's visit to the Ghent[2]), art (such as the STAM painting of Ferdinand III[3] and the painting of Maria Theresia as a gift to the city of Ghent), place and street names (e.g. Koningin Maria Hendrikaplein and Jozef II-straat in Ghent), and memories of foreign academics/researchers (that probably are local heroes in their own region). The project results will give a better view on the diversity and shared histories and on multiple connections between our countries to de-nationalize people's views. An overview of the EURECA methodology is shown in Figure 1.

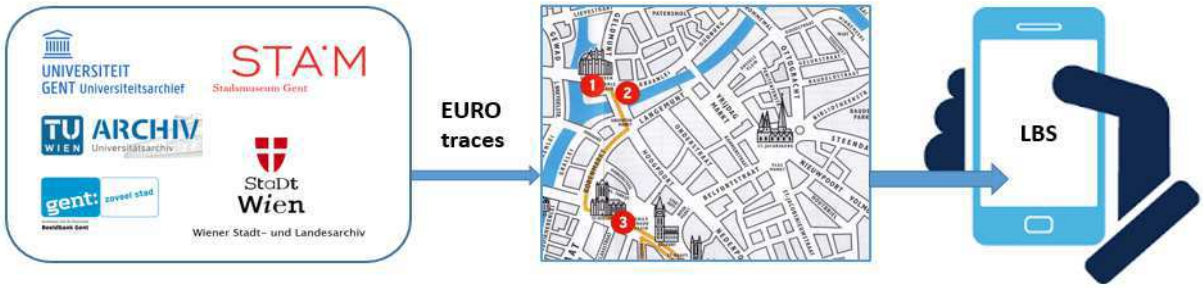

Figure 1: EURECA methodology for the enrichment, mapping and querying of European traces in city archives.

The current collections of the city archives, however, have some technical and usability bottlenecks which make it difficult to easily find European traces and map them to a location on the map. The current metadata scope of photo archives, for example, is too narrow and too high-level to allow easy and adequate exploration of the collection data (e.g. to find regional traces or similar images). The main goal of the EURECA project is to address these issues, and increase the searchability of the collection items (mainly focusing on the location and traces aspect). First of all, geographic entity recognition (GER), which can be run on textual metadata and image content, will help in extending and linking the existing collection items and facilitate spatial collection mapping for interactive querying. Secondly, expert tagging micro tasks and crowdsourced techniques will be used to address the missing (meta)data problems and perform validations on the automatically generated location and traces (meta)data. Furthermore, a spatiotemporal dashboard will be developed to study/analyze the spatial and temporal features and evolution of the traces and compare them to cultural heritage-related social media footprints of what tourists of a particular region of origin currently visit in the city. Finally, we will develop guidelines and methodologies for the creation of LBS and visualizations of region-specific POIs. Within this paper, however, we only focus on the preliminary results of the project, which is the clustering/recognition of handwritten text for GER and the analysis/visualization of social media footprints.

[2] http://www.gandante.be/rococo-in-gent-met-een-snuifje-mozart/
[3] http://stamgent.be/nl_be/collectie/kunstwerken/00716

The remainder of the paper is organized as follows. Section 2 focuses on the spatio-temporal analysis and recognition of geographic entities in handwritten student registers of Ghent University. Next, Section 3 discusses the social media based detection of Areas of Interest (AoI's) that represent popular places for visitors from specific regions of origin. These AOIs are characterized by the most common terms extracted from the semantics contained in GeoSocial Media (GSM) posts. Finally, Section 4 lists the conclusions and points out directions for future work.

2. Improving HTR predictions for geographic entities

Archive collections contain vast amounts of historical documents, with a digital scan of the most important documents. Transcribing all these documents so that these can be searched and indexed requires an enormous amount of work for large collections. State of the art handwritten text recognition (HTR) solutions can aid this process by automatically transcribing these documents, however, these often contain a large number of errors. Manually correcting all these errors can be just as demanding as transcribing the full documents. In order for a HTR solution to work well enough, the model needs to be trained with hundreds of correctly labeled pages which are often not available.

In this work, we will analyze how these out of the box HTR solutions can be improved in an unsupervised way. Our dataset is the Ghent university student register archive, which is structured in a table (as shown below). The full collection contains records spanning over 100 years, which means that there were several writers, each with a different writing style, making it difficult to accurately predict the transcriptions. The table contains each student's ID, name, birthplace, age, sex, faculty of enrollment and field of study. The collection contains both Dutch and French words. We have focused mainly on the birthplace column (~ geographic entity), as this makes it possible to link this data to a map and query it based on a particular region.

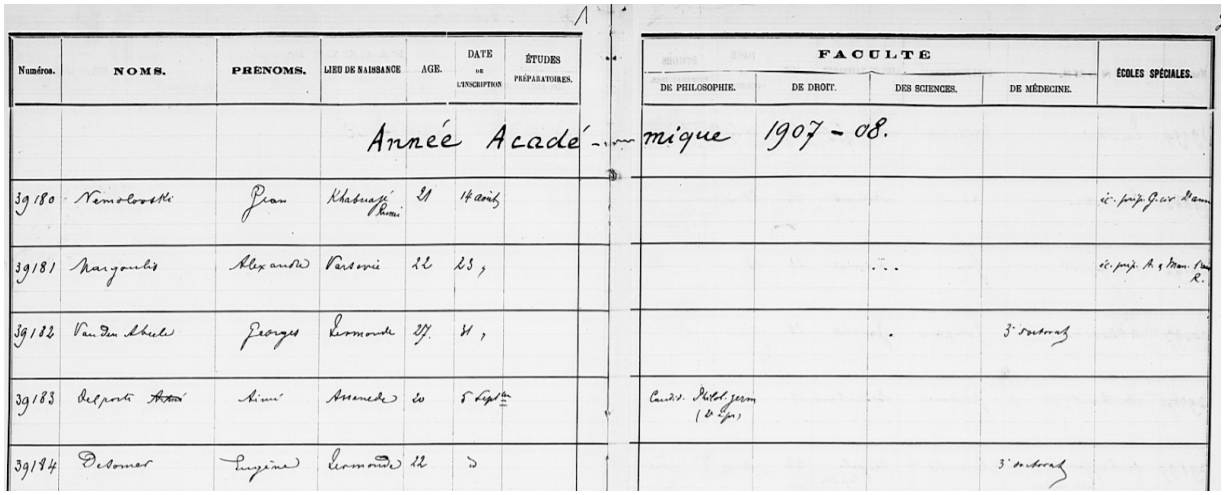

Figure 2: Handwritten student registers of Ghent University structured in table format.

The first step of any HTR workflow is to extract the text regions of each page. In a full-text page, this is done through line detection (Gruning, 2018), but as we will focus on tabular data, the lines of the table are extracted through computer vision techniques and the bounding box of each cell is cut out. Once the data is collected, these are preprocessed. Noise is removed, the text is cut out and all images are resized to the same dimensions. Then, the images are transcribed with a simple, pretrained HTR model[4] on the IAM dataset[5]. These transcriptions form the baseline predictions, which we aim to improve in an unsupervised fashion.

The main idea behind our approach is the fact that place names are often repeated and that each of these representations will share some visual similarity. We can leverage this similarity in order to improve HTR results in a post-processing step. In order to validate this assumption, a small subset of 30 pages was manually labeled. The pretrained HTR resulted in a character error rate (CER) of 53.3 % and a word error rate (WER) of 95.9 %, making it difficult to match the transcription to the right label. Our solution for this is simple, given a query image, the k most similar images are collected, then the HTR predictions of the query image and of all of these similar images are used to calculate the most probable word length and letter at every position for the query image. As the HTR model will often predict one or two letters wrong, this averaging will cancel out these errors. In order to find the most similar images of a given image, a word spotting approach similar to (Kovalchuk, 2014) was used that scored a top-3 mean average precision (mAP) of 0.797. Two different experiments were done, the first changes the transcription of each image with the average transcription of all the images containing the same text, this will give the best-case outcome assuming that our word spotting algorithm works perfectly. The second experiment is similar, but the transcription is changed with an average of 5 randomly chosen images with the same text. The first experiment improved the CER to 30 % (20% net gain), the second one was repeated multiple times due to the random selection and improved the CER from 31.4 % to 50 %. These experiments show that most of the HTR errors are canceled out by statistical averaging the transcriptions, but assume that we know a priori how many images contain the same text (some place names have less than 5 occurrences) and that our visual similarity approach works perfectly.

Figure 3: Visually similar words of 4 images with the text ‘Doornik’, the matches for the third query image are not correct.

[4] https://github.com/githubharald/SimpleHTR
[5] http://www.fki.inf.unibe.ch/databases/iam-handwriting-database

Next, an unsupervised approach was used: given a query image, its 3 most similar images are found through visual similarity and its transcription is changed to the most probable transcription of these 4. This approach only improved the CER by 1 %, showing the importance of the visual similarity algorithm. This was expected, as the transcriptions of non-relevant words will induce additional errors. For very large collections, where each unique word has a lot of occurrences however, these non-relevant words will have less impact, as the transcriptions can be averaged for 20 or 50 similar words. The gain in prediction accuracy can also be minimal because the starting HTR predictions are too bad, and taking multiple bad predictions will not cancel out small mistakes. To further validate this approach, it was repeated on the GW dataset[6]. The pretrained model scored very poorly (> 80% CER), so it was retrained on the GW dataset, using only 20% of the dataset as training data. The HTR model achieved a CER of 22.6 % and a WER of 41.0 %, validated only on the words with 3 or more occurrences (~ 3000 words). The visual similarity algorithm scored a top-3 mAP of 0.787.

The same approach as above was applied, but the 5 most visually similar images were taken and the HTR prediction was only swapped to the most probable prediction if the average CER between all predictions was below a certain threshold (to reduce the effect of errors in the visual similarity algorithm). This improved the CER to 19.1 % and the WER to 34.4 % (~200 more words correctly predicted), validating the effectiveness of this simple approach. The algorithm made 266 correct swaps (swap prediction with the correct label), 373 good swaps (swap that reduces the CER) and 148 bad swaps. The algorithm changed 787 predictions (26 % of total), from which 71.5 % were beneficial. Note that no labeled data was used for this improvement, having a portion of correctly labeled data would only further reduce the error rate, as these images can be given a larger weight during the prediction averaging. This approach can be further improved by using a word list of historical place names and swapping the prediction with the closest label. Furthermore, this approach can be used to augment traditional HTR workflows, as these look at each word image independently, while this approach uses the entire dataset as a post-processing step to cancel out a lot of the small mistakes that HTR model makes.

Another possible application of the proposed methodology/solution is a smart incremental labeling tool that can be made for documents containing a lot of similar data (e.g. tables & forms, or even full text like GW dataset). The tool calculates visual features and performs an HTR prediction for all images, then selects the best 10-20 matches for each image based on visual similarity. The tool then gives the user a query image and its best 10-20 visual matches, with the average HTR prediction of these. The user selects which images are not the same as the query image, and enters the correct label for all the similar ones once. This new input can be used to increase the confidence of the next predictions of similar images. The more labeled data, the more accurate the predictions. This allows for faster labeling, instead of typing each label on each word, the user can label 10-20 images at a time. This speedup will decrease when the user has labeled most images, as only the non-frequently occurring words will remain.

[6] http://www.fki.inf.unibe.ch/databases/iam-historical-document-database/washington-database

As a next step, we will investigate if we can estimate for each query word, how many words with the same transcription there are in the dataset by looking at the distribution of the distances between the visual features of each word. That way, we can dynamically adjust how many predictions of visually similar images will be considered. The mistakes which the HTR model and our algorithm make will be analyzed in order to potentially apply a non-uniform weighting scheme to letter probabilities; impossible and non-frequent letter combinations will also be filtered out or assigned a lower weight.

4. Social media based Areas of Interest (AoI) detection

The first phase of the AoI detection involves the data collection from Flickr via two of its APIs. Metadata of each uploaded picture such as the photo owner, uploading date, and geolocation was retrieved. In a second process, another API was used to obtain the user name, location (user manually-provided) and other attributes. This location attribute had to be processed because of the heterogeneity of the data provided. The identified places had to be matched with the GeoNames gazetteer[7] to determine the country these places are in. The data retrieved covered a squared area of 68 Mill. km2 representing a huge area around the continental Europe. In order to determine the Region of Origin (RoO) of each user the first source of information is the self-reported location included in her profile. Unfortunately, this information is often missing or can be simply false. For the majority of the users, the origin had to be inferred by some kind of method. A simple method based on previous works on home determination from user's GSM data (Li & Goodchild, 2012; Paldino et al., 2015) was developed and tested. To identify a country as RoO location for a user, all the pictures uploaded by her in all the countries of analysis were considered. Among the countries in which the user had uploaded pictures for a period greater than 6 months, the one with the highest number of total pictures was selected as origin for the user. This information was used to determine the RoO for those users without location information in their profiles.

We used the geolocation of the photos (as points) for visualization. A continuous raster surface was generated from these points using Kernel Density Estimation (KDE) (Grothe & Schaab, 2009). These raster are heatmaps that represent areas of high concentration of pictures. The heatmaps represent a footprint of the visitors in the city. Thus, the areas more visited by tourists from a specific origin were visible and an analysis of their temporal evolution will be possible. The continuous surfaces built with KDE are very well suited for the task of determining vague areas open enough for further POIs identification in EURECA. Figure 4 shows examples of footprints in Vienna and Ghent.

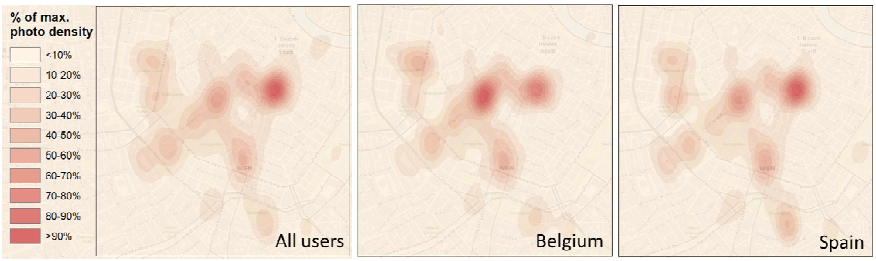

Figure 4: KDE footprints for Vienna: All users, RoO Belgium & RoO Spain

[7] www.geonames.org

Elements located throughout a network are often analyzed with spatial methods assuming Euclidean distance. Nevertheless, Euclidean distances and their equivalent short-path distances are significantly different (Okabe & Sugihara, 2012). Network Kernel Density Estimation (NKDE) has been used in previous works (Delso et al., 2018; Okabe et al., 2006) for the estimation of the density of points on a network. This is partly the case when studying the distribution of pictures along a city because users move through the street network. Hence, we applied NKDE in order to obtain also footprints along the street network. In further work, this will allow us to compare both types of footprints, identify streets of interest and estimate intensity of touristic usage along the network. The ArcGIS toolbox SANET (Okabe et al., 2018) was used for the NKDE in Vienna and Ghent. The footprints (shown below) revealed the most preferred places for specific RoO. Furthermore, all the footprints were compared through spatial analysis. Using map algebra (Tomlin, 1994), we obtained areas of common interest for specific nations of origin.

Figure 5: NKDE footprints for Vienna: All users, RoO Belgium & RoO Spain

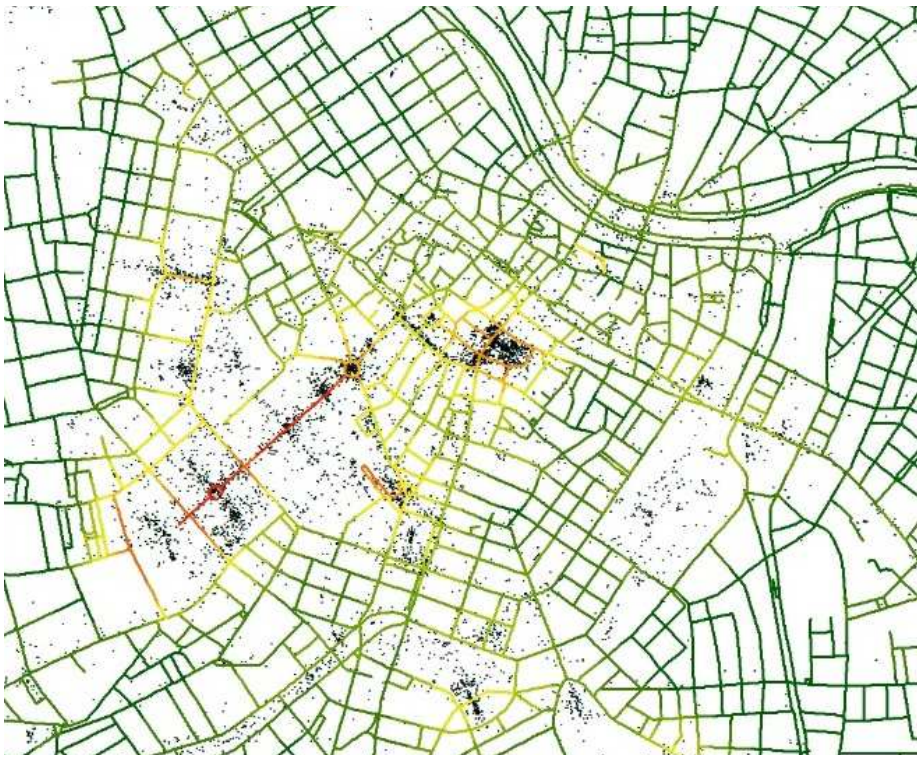

Figure 6: NKDE on Flickr pictures for the center of Vienna

A study area was selected to perform an additional analysis of the RoO footprints. This place is located in the city of Vienna (Austria) and consist in a 1km diameter circle including the *Karlsplatz* area. This place is a typical touristic hotspot in the city that offer a variety of cultural heritage POIs. For this analysis we used NLP techniques to extract the terms contained in the tags associated with the photos, along with their frequencies of appearance. Using common tools in social media data mining, we generated a graph whose nodes represented the terms and whose edges reflected relations between terms. The size of the nodes is proportional to the total frequency of appearance and the thickness of the edges is proportional to the frequency of co-appearance in photos. Then we were able to visualize the most used terms with a glimpse on how often they are used in the area and how commonly they are used together. The added value lies on the characterization of the space by the most often semantics generated by users depending on their RoO. This provides some kind of a DNA of the space.

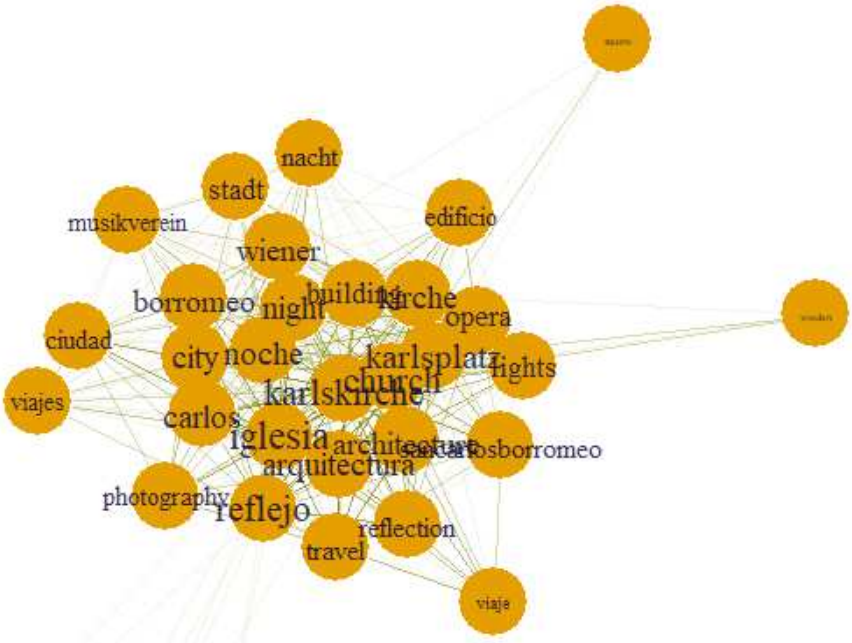

Figure 7: Word graph for RoO Spain in Karlsplatz

To summarize, the final number of points retrieved was about 66 million and covered a period (2004-2018) representing Flickr photos from 62 countries. Initial research was done with a selection of 2 European cities and countries: Ghent (Belgium) and Vienna (Austria). The footprints generated showed differences among visitors according to their country of origin. It is possible identifying areas of special interest for some nations. The KDE footprints discovered diffuse continuous areas whereas the NKDE footprints showed genuine patterns of street segments more used by visitors from certain nations. The word graphs allowed identifying terms related to specific nationalities. This was clear with words in national languages of the RoO. Also specific cultural heritage topics are more frequent when considering some nations, though further research is needed in order to establish more clear relations to the POIs present in the area. Further work will be necessary to optimize the different analyses and draw additional conclusions.

5. Conclusions

The proposed EURECA project aims to establish connections and collaboration among Flemish and other European cultural heritage archives. By enriching their collections with European traces, we provide them new types of relationships that they can use to link their collection items/datasets and perform cross-collection analysis/studies. Both the cultural heritage traces and POI (meta)data can be used as input to perform new fundamental research and applied studies, but also to facilitate the exploitation of the collections to a broader public and attract new groups of cultural heritage consumers, i.e., the EURECA project will increase awareness and access to cultural heritage. Furthermore, the European traces will strengthen the Europe in Europe feeling and reveal the European DNA of our cities. Finally, our dynamic routing will allow end-users to easily consume cultural heritage whilst exploring the city/region and crowdsourced micro-tasks (which are part of future work) will attract their attention.

References

Delso, J., Martín, B., & Ortega, E. (2018). A new procedure using network analysis and kernel density estimations to evaluate the effect of urban configurations on pedestrian mobility. The case study of Vitoria –Gasteiz. *Journal of Transport Geography*.

Grothe, C., & Schaab, J. (2009). Automated footprint generation from geotags with kernel density estimation and support vector machines. *Spatial Cognition and Computation*, 9(3), 195–211.

Grüning, T., Leifert, G., Strauß, T.. & Labahn, R. (2018). A Two-Stage Method for Text Line Detection in Historical Documents.

Kovalchuk, A., Wolf L., & Dershowitz, N. (2014). A Simple and Fast Word Spotting Method. *14th International Conference on Frontiers in Handwriting Recognition*, pp. 3-8.

Li, L., & Goodchild, M. F. (2012). Constructing places from spatial footprints. In Proceedings of the 1st ACM SIGSPATIAL International *Workshop on Crowdsourced and Volunteered Geographic Information - GEOCROWD '12* (p. 15).

Okabe, A., Okunuki, K. I., & Shiode, S. (2006). SANET: A toolbox for spatial analysis on a network. *Geographical Analysis*.

Okabe, A., Okunuki, K., & SANET Team. (2018). SANET. A Spatial Analysis along Networks (Ver.4.1). Tokyo, Japan.

Paldino, S., Bojic, I., Sobolevsky, S., Ratti, C., & González, M. C. (2015). Urban magnetism through the lens of geo-tagged photography. *EPJ Data Science*, 4(1), 1–17.

Tomlin, C. D. (1994). Map algebra: one perspective. *Landscape and Urban Planning*.